International Journal of Advanced Science and Technology
Vol.113 (2018), pp.61-70
http://dx.doi.org/10.14257/ijast.2018.113.07

Hybrid Processing Technique to Gender Inference in Social Network Services

Ch Sudhakar[1], A Sravani[1], N. Thirupathi Rao[1], Debnath Bhattacharyya[1] and Tai-Hoon Kim[2*]

[1]Department of Computer Science & Engineering
Vignan's Institute of Information Technology (A)
Visakhapatnam, AP, India
[2]Sungshin Women's University, Bomun-ro 34da-gil, Seongbuk-gu, Seoul, Korea
sudhakarcheetirala@gmail.com sravani61sravz@gmail.com
[1]nakkathiru@gmail.com, [1]debnathb@gmail.com,
[2]taihoonn@daum.net*

Abstract

Now a day's most of them are using social networking services (SNS) on regular basis. Twitter is that the one amongst the foremost common SNS wherever users post news, messages and conjointly exploitation for online public polling. As we all know Twitter already introduced online polling wherever users selecting appropriate suggestion's, leaders etc. Profile reasoning of social networking services users is effective for on-line polling, public, and personal selling and subject matter. Twitter profiler gender reasoning is examined by exploitation deep learning, image process, NLP. Earlier text process used for gender reasoning and conjointly exploitation image recognition techniques. Currently this paper introducing the hybrid techniques of image and text algorithms area unit exploitation for distinguishing gender reasoning. The attribute is provided from text mining algorithms and from image process techniques area unit combined to profiler gender reasoning.

Keywords: *SNS, Twitter, Text Mining, Image Recognition*

1. Introduction

Now a day's most of them are the use of social networking services (SNS) for share and changing views, guidelines, evaluations, information on several matters. Maximum of the corporations are encouraging to apply social networking services (SNS) for enhancing great and quantity in their products and services like flipkart, amazon and so forth. The primary problem is the SNS profile having their name, gender, age, a residence which is not openly to be had, however such facts is rather important for advertising. Then only the usage of this novel hybrid set of rules for identifying infer demographic facts of unknown customers. The summation of all consequences from the textual content mining and photograph processing algorithms to finding inference of each user of SNS.

In this paper, we proposed novel technique the aggregate of textual content and photo algorithms concurrently to locating SNS consumer gender inference. On this paper, we tested eastern twitter user's text and image facts. The novel set of rules takes tweets as input and gives output as a gender opportunity score of the consumer. The algorithm works in two important steps. Step (1) using processors are to locate gender probability ratings. Step (2) merging two chance rankings are based totally at the specific ratio. This result will supply records concerning social community carrier user gender inference.

Received (October 18, 2017), Review Result (January 10, 2018), Accepted (January 17, 2018)

[*] Corresponding Author

ISSN: 2005-4238 IJAST
Copyright ⓒ 2018 SERSC Australia

This paper is divided into five phases. Phase 2 represents the facts approximately textual content mining and picture recognition. Phase 3 explains regarding proposed novel hybrid set of rules. Phase 4 will explains effects of the proposed algorithm and locating out gender inference. Phase 5 discusses and future strategies and summarize this paper.

2. Related Work

There are many existing areas are available for identifying gender inference. They are text mining, image processing and machine learning *etc*. Many algorithms are introduced like Support Vector Machine. Also many classifying algorithms are used for text mining. Earlier examined demographic feature inference for SNS users based on machine learning. Sharing text and images in a Social Network Services, the user information is in virtual cloud. Twitter profile data as trained data for machine algorithms. The following processors are using on twitter data and get required information.

Text Processing:

The textual content processing is the way of retrieve the predictable information through the usage of mining classifiers. Text processor can handle the textual content facts that are accrued from person tweets in a SNS. Text processor will take textual content tweets as an input and gives gender possibility rating of user as an output after performing of text classifier. The mining algorithm aid vector system is used on textual content processing for getting chance scores of gender inference.

The content based strategy methodology is the accompanying.

Stage 1-1 Tokenization is finished utilizing Kuromoji (http://www.atilika.org), a Japanese morphological analyzer. Along these lines, the unigram is acquired. At that point, the pack of words highlight is extricated from the unigram.

Stage 1-2 The SVM gets the sack of-words include as info. The male likelihood score is acquired utilizing SVM. At that point, the female likelihood score is ascertained utilizing Equation 1.

$$scoremale + scorefemale = 1$$

Image Processing:

Another step of finding gender inference is image processor which is performed on two steps. In the first step, the annotating images by an image processing techniques. In the second step, find and measuring probability score of gender as male or female according to the first step at the user level.

The picture based technique strategies are the following.

Stage 2-1 Probability score of pictures for subcategories is acquired utilizing the CNN model.

Stage 2-2 The score of every client is acquired by averaging the likelihood score of pictures that the client posts since numerous clients posted more than one picture.

3. Proposed Hybrid Novel Algorithm

Our proposed novel hybrid algorithm for combination of text mining and image recognition is proven in fig 1. Take input as a thousand tweets which are posted by way of a profiler are labeled into text and picture statistics. Each man or woman processor analyses the text and photo records one by one.

The hybrid novel based strategy grouping scores related to content-based and picture based technique yield and gauges the gender inference of individuals who posted content and pictures. We utilized calculated relapse.

In stage 3-1, the male likelihood score is gotten utilizing strategic relapse. At that point, the female likelihood score is figured utilizing Equation 1 in a similar way as that introduced in stage 1-2.

The preparation procedure for calculated relapse includes two phases. In the primary stage, the text-based what's more, picture-based strategy is prepared to get preparing information for the half and half based strategy.

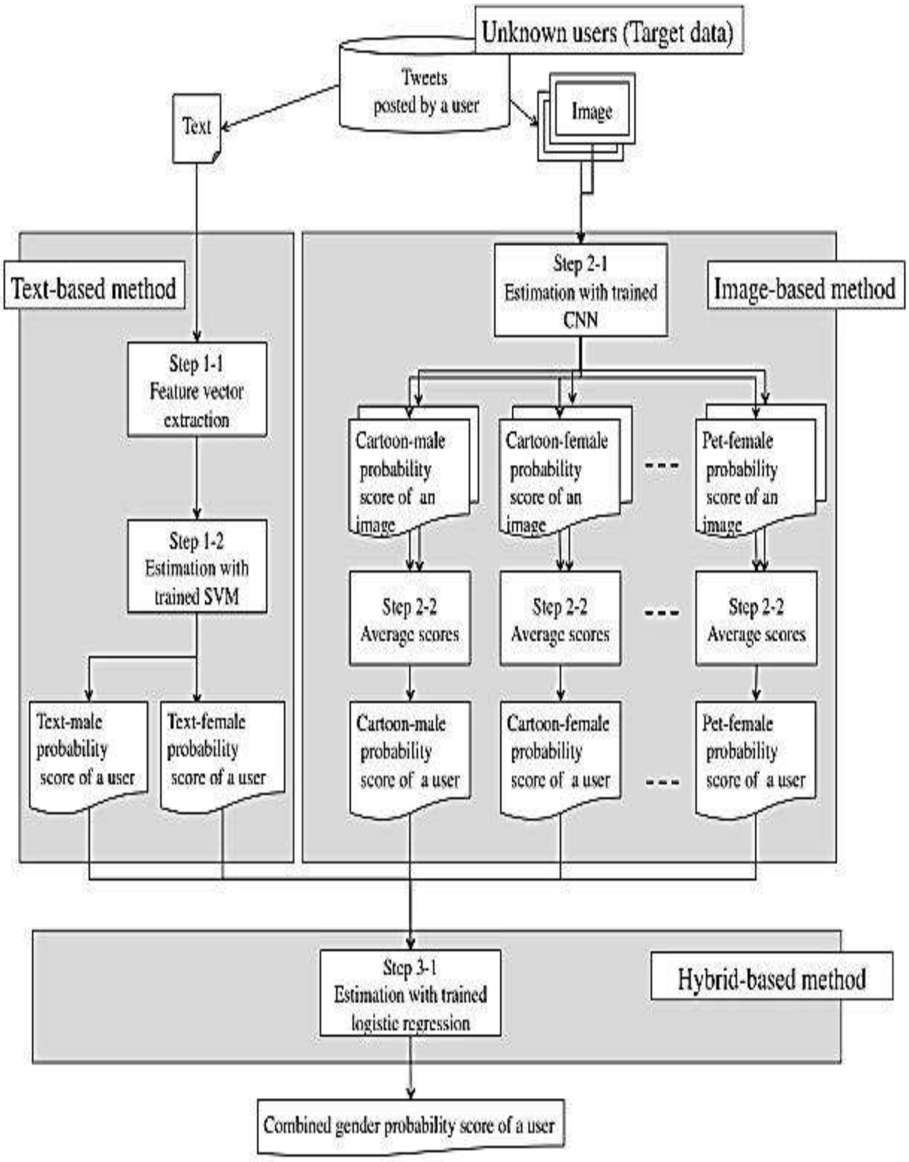

Unknown users (Target data)
Tweets posted by a user
Text
Image
Text-based method
Image-based method
Step 2-1 Estimation with trained CNN
Step 1-1 Feature vector extraction
Cartoon-male probability score of an image
Cartoon-female probability score of an image
Pet-female probability score of an image
Step 1-2 Estimation with trained SVM
Step 2-2 Average scores
Step 2-2 Average scores
Step 2-2 Average scores
Text-male probability score of a user
Text-female probability score of a user
Cartoon-male probability score of a user
Cartoon-female probability score of a user
Pet-female probability score of a user
Step 3-1 Estimation with trained logistic regression
Hybrid-based method
Combined gender probability score of a user

Figure 1. Combination of Text and Image Processing

The summation of probability scores from each individual processor and finally gives probability score for gender inference.

Copyright ⓒ 2018 SERSC Australia

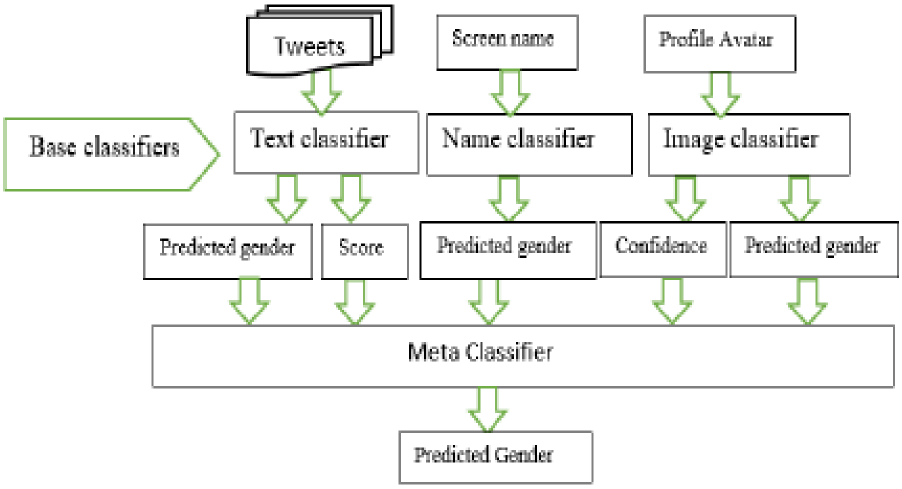


Figure 2. System Architecture for Finding Gender Inference

SVM Classifier Method:

To combine these two results which are derived from text and image processing by using Scoretext and Score image.

$$Score_{combined} = Score_{text} \times \alpha + Score_{image} \times (1-\alpha)$$

To find the attribute value for n-th top word W, for a specific user, we can evaluate to the more appeared occurrences of words of tweets in a SNS. n-th top word. The most appeared word of tweets which are posted by the user. By identifying these most occurred words in a tweets, word W appearing in a post by the formula

Score (Words) =ul1 (Words)-ul2 (Words)

Where ul1 represent as the frequency of words appeared in user lable1. Like that ul2 represent as the frequency of words appeared in user lable12.

$$\frac{number\ of\ occurrences\ of\ words}{word\ count\ of\ profilers}$$

Logistic regression:

We select influential predictor variables based on Akaike's Information Criteria (AIC) (Akaike, 1973), which represents adaptability of built models. The variables selection aims to minimize AIC. Here, we adopt backward elimination method for the selection.

Random forest:

This method measures mean decrease Gini index of each predictor variable. The Gini index reveals the extent of deviation of a classification result. The smaller the deviation, the better the classification result. Therefore, variables having larger mean decrease Gini index are regarded as better predictors. The number of variables for selecting the best performance is based on the mean decrease Gini index.

Copyright ⓒ 2018 SERSC Australia

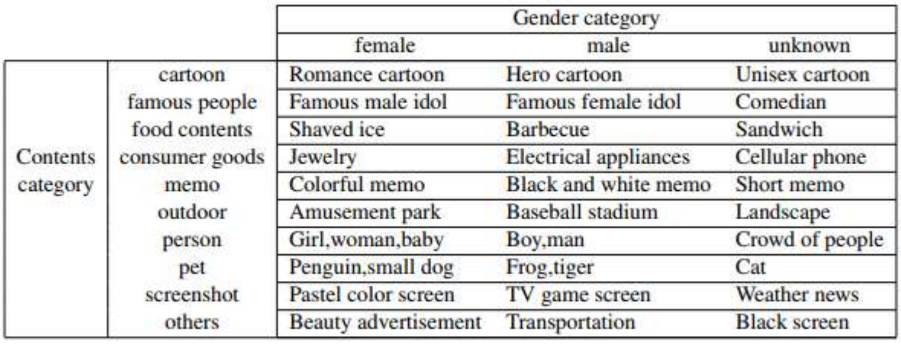
SVM:

Instead of selecting influential variables, SVM tunes two internal parameters: cost and gamma, using grid search. Cost determines the extent of wrongly classified instances, and gamma represents boundary simplicity. An RBF Gaussian kernel is used for base conversion.

4. Experimental Data & Analysis

Within the proposed gadget we've got used facts set related to extraordinary classes of pics which are associated with real time activities consisting of cartoons, famous user's images and food merchandise and so forth. Sub categories with gender class also shown beneath. The subsequent desk includes typical contents of sub categories obtained using image level annotation.

Table 1. Sub Category Composed with the Combination of the Gender and Contents Category



		Gender category		
		female	male	unknown
Contents category	cartoon	Romance cartoon	Hero cartoon	Unisex cartoon
	famous people	Famous male idol	Famous female idol	Comedian
	food contents	Shaved ice	Barbecue	Sandwich
	consumer goods	Jewelry	Electrical appliances	Cellular phone
	memo	Colorful memo	Black and white memo	Short memo
	outdoor	Amusement park	Baseball stadium	Landscape
	person	Girl,woman,baby	Boy,man	Crowd of people
	pet	Penguin,small dog	Frog,tiger	Cat
	screenshot	Pastel color screen	TV game screen	Weather news
	others	Beauty advertisement	Transportation	Black screen

4.1. Datasets

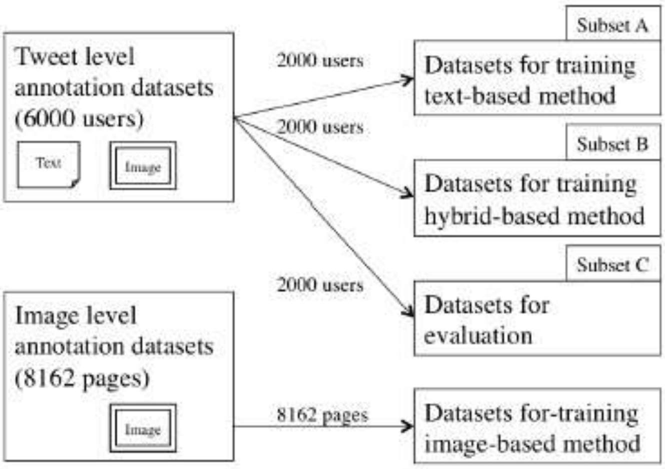
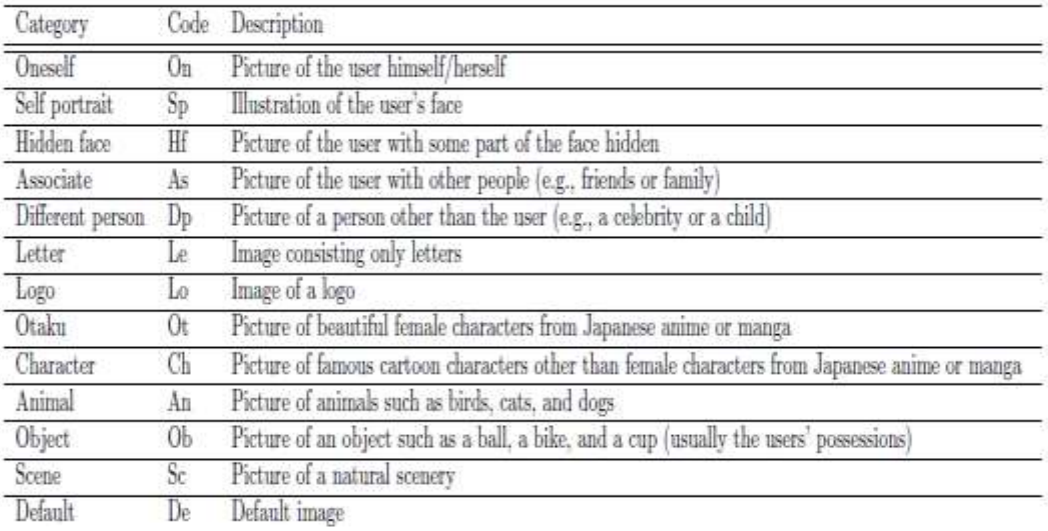

Tweet level annotation datasets (6000 users)
Text
Image
Image level annotation datasets (8162 pages)
Image
2000 users
2000 users
2000 users
8162 pages
Subset A
Datasets for training text-based method
Subset B
Datasets for training hybrid-based method
Subset C
Datasets for evaluation
Datasets for-training image-based method

Figure 3. Datasets for Experimenting

In Figure 3 shows, the tweet level explanation information was part up into three subsets. Subset A was utilized for preparing the content based strategy. Subset B was utilized for preparing the crossover based technique. Subset C was utilized for assessment. Picture level comment information was utilized for the preparation picture based technique.

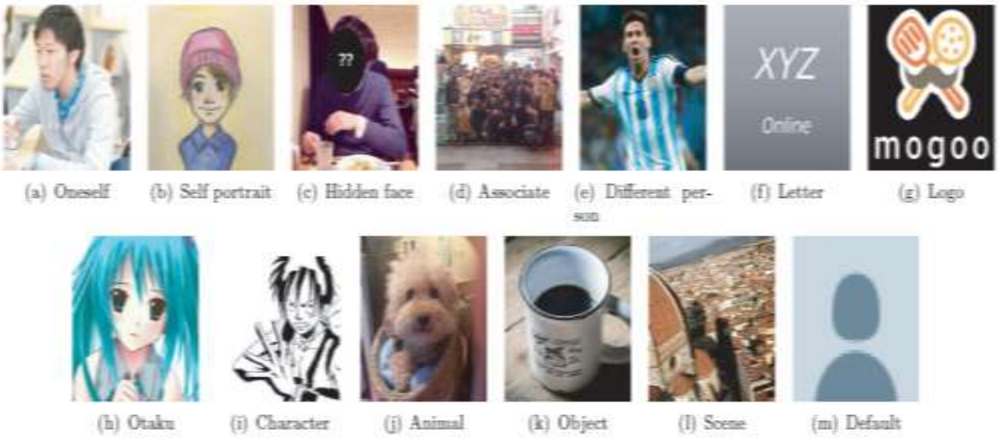
Copyright ⓒ 2018 SERSC Australia

Table 2. The Categories, Codes and Descriptions of Profile Images



Category	Code	Description
Oneself	On	Picture of the user himself/herself
Self portrait	Sp	Illustration of the user's face
Hidden face	Hf	Picture of the user with some part of the face hidden
Associate	As	Picture of the user with other people (e.g., friends or family)
Different person	Dp	Picture of a person other than the user (e.g., a celebrity or a child)
Letter	Le	Image consisting only letters
Logo	Lo	Image of a logo
Otaku	Ot	Picture of beautiful female characters from Japanese anime or manga
Character	Ch	Picture of famous cartoon characters other than female characters from Japanese anime or manga
Animal	An	Picture of animals such as birds, cats, and dogs
Object	Ob	Picture of an object such as a ball, a bike, and a cup (usually the users' possessions)
Scene	Sc	Picture of a natural scenery
Default	De	Default image

Table 2 Explaining records brief description of each of the 13 classifications. Additionally, we demonstrate test profile pictures relating to these classes in Figure 1. In figures and tables seeming later in this paper, we utilize the codes appeared in the middle segment of Table 1 to speak to the recorded classes (*e.g.*, "On" remains for oneself). In the following subsections, we disclose our client trials to make these strides.



(a) Oneself (b) Self portrait (c) Hidden face (d) Associate (e) Different person (f) Letter (g) Logo

(h) Otaku (i) Character (j) Animal (k) Object (l) Scene (m) Default

Figure 1. Sample Profile Images of the 13 Categories

We establish 13 categories of objects observed in users' profile images: “oneself", “self-portrait", “hidden face", “associate", “different person", “letter", “logo", “otaku", “character", “animal", “object", “scene", and “default".

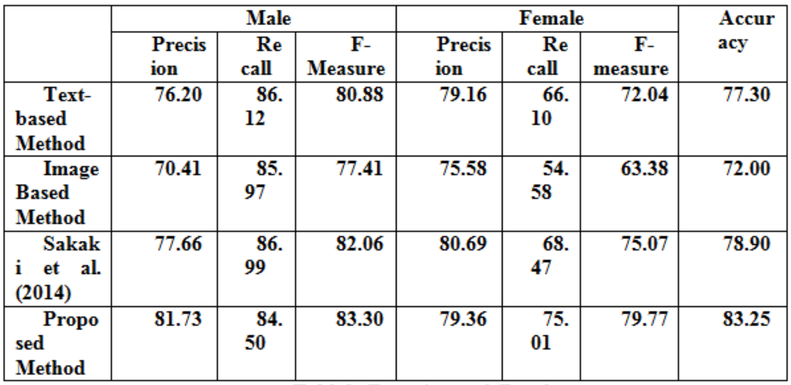
4.2. Experimental Setup and Results

Table 2 indicates the precision, recollect, f-degree, and accuracy of our proposed machine. The accuracy of our proposed method carried out 83.25 percentages which is 5.95 percent higher than that of the textual content-based totally technique, 11.25 pt better than that of the picture-based totally approach, and 4.35 pt higher than that of the approach defined by sakaki *et al.*, (2014). Mainly, the girl f-measure related to our proposed approach carried out seventy nine.77 percentage, which is 7.seventy

Copyright ⓒ 2018 SERSC Australia

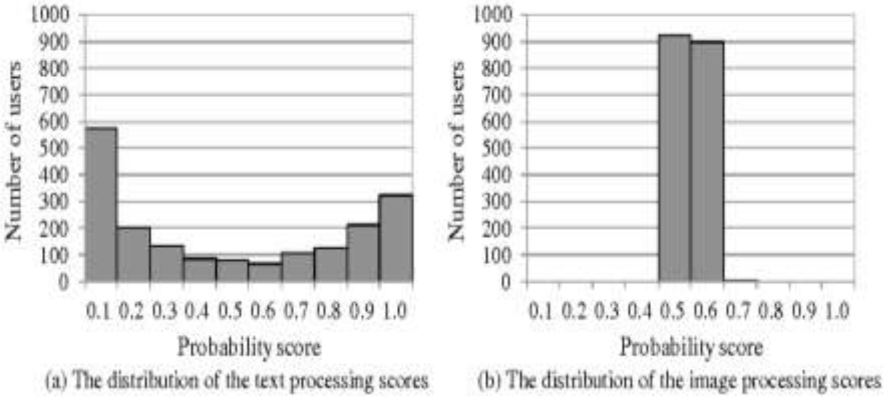
three pt higher than that of the text based totally approach, 7.seventy seven pt higher than that of the picture primarily based technique, and 4.7 pt better than that of the approach defined via sakaki *et al.*, (2014). We conducted a binomial take a look at to assess our proposed approach and the technique described by way of sakaki et al. (2014). Results confirmed that the p price is zero.0028, which indicates that the outcomes received the usage of our approach is considerably higher than the ones received the use of the prevailing aggregate based totally method.

Table 2. Experimental Results



	Male			Female			Accuracy
	Precision	Recall	F-Measure	Precision	Recall	F-measure	
Text-based Method	76.20	86.12	80.88	79.16	66.10	72.04	77.30
Image Based Method	70.41	85.97	77.41	75.58	54.58	63.38	72.00
Sakaki et al. (2014)	77.66	86.99	82.06	80.69	68.47	75.07	78.90
Proposed Method	81.73	84.50	83.30	79.36	75.01	79.77	83.25

The following graph is drawn by way of taking statistics units on x-axis and logistic regression weights on y-axis. Man or woman t denotes weights w.r.t textual content based method and person i denotes weights w.r.t picture primarily based technique. It shows the logistic regression weights. From this determine, we located that the weights for lady customers were all nice; the weights for male customers have been almost all negative. The weights for unknown were almost 0, which suggests that the possibility scores of text-based totally and photograph-primarily based techniques are not competing.

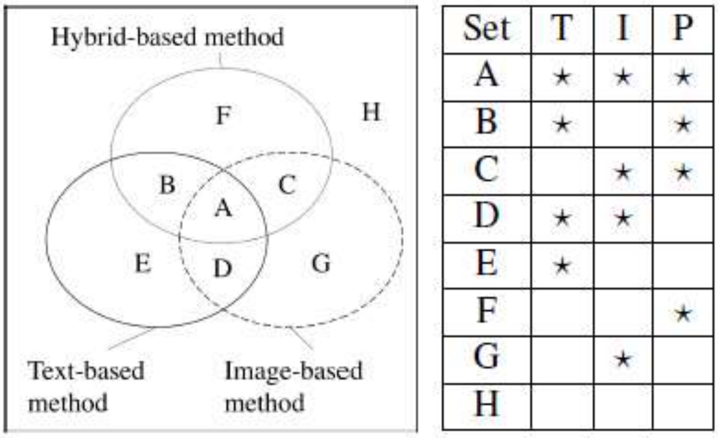


(a) The distribution of the text processing scores (b) The distribution of the image processing scores

Figure 4. Distribution of the Probability Scores

4.3. Hybrid based Method:

This section exhibits an exchange of the adequacy of the blend of the content based and picture based techniques. It addresses the contrast between the model proposed by Sakaki *et al.*, (2014) and our proposed technique. At long last, the utilizations of strategic relapse weights of the joined sources are examined.



Set	T	I	P
A	⋆	⋆	⋆
B	⋆		⋆
C		⋆	⋆
D	⋆	⋆	
E	⋆		
F			⋆
G		⋆	
H			

Figure 5. Novel Hybrid Based Technique

The above Figure 5 represents T means content-based technique. I signify picture based strategy. P means novel hybrid-based strategy. "⋆" signifies the set of clients whose gender orientations were gathered accurately utilizing each strategy. Each circle of the Venn graph specifies to an set of clients whose gender orientation was induced accurately utilizing a strategy. The union of A, B, D, and E specifies to clients whose gender orientation was induced accurately utilizing the content based technique. B tells to clients whose gender orientation was gathered effectively by the content based and the hybrid-based technique, however was misinterpreted utilizing the image-based technique. Maximum of the corporations are encouraging to apply social networking services (SNS) for enhancing great and quantity in their products and services like flipkart, amazon and so forth. The primary problem is the SNS profile having their name, gender, age, a residence which is not openly to be had, however such facts is rather important for advertising. Text processor can handle the textual content facts that are accrued from person tweets in a SNS. Text processor will take textual content tweets as an input and gives gender possibility rating of user as an output after performing of text classifier. The mining algorithm aid vector system is used on textual content processing for getting chance scores of gender inference.

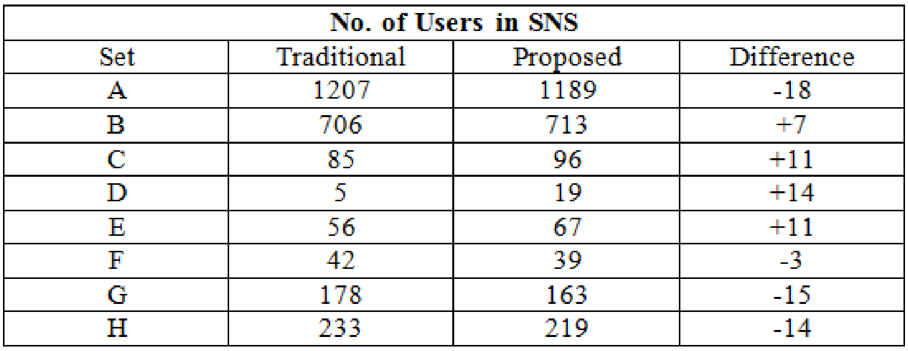
Then only the usage of this novel hybrid set of rules for identifying infer demographic facts of unknown customers.

The below Table 3 shows an difference of methods of traditional and proposed techniques of users in SNS. We might want to look at C, D, E, and F particularly to survey the distinction in the execution between the content based and the hybrid-based strategy. C and F incorporate clients whose separate gender orientations were gathered effectively utilizing the half and half based technique, yet misconceived utilizing the content-based strategy. The weights for unknown were almost 0, which suggests that the possibility scores of text-based totally and photograph-primarily based techniques are not competing. D and E incorporate clients whose separate gender orientations were gathered effectively utilizing the content based strategy, however, misinterpreted utilizing the half and half based technique. Text processor can handle the textual content facts that are accrued from person tweets in a SNS. Text processor will take textual content tweets as an input and gives gender possibility rating of user as an output after performing of text classifier. The mining algorithm aid vector system is used on textual content processing for getting chance scores of gender inference.

Each circle of the Venn graph specifies to an set of clients whose gender orientation was induced accurately utilizing a strategy. Here we talk about the

Copyright © 2018 SERSC Australia

outcomes gotten utilizing the proposed strategy, which are appeared in the third section.



No. of Users in SNS			
Set	Traditional	Proposed	Difference
A	1207	1189	-18
B	706	713	+7
C	85	96	+11
D	5	19	+14
E	56	67	+11
F	42	39	-3
G	178	163	-15
H	233	219	-14

Table 3. No of Users Included in each Traditional vs. Proposed Methods

5. Conclusion

The fundamental advantage of our proposed framework is that highlights gave from a content classifier and from a picture classifier are consolidated properly to identify male or female sexual orientation utilizing strategic relapse. Test comes about demonstrated that our approach accomplished exactness of 83.25 percent, which was 5.95 pt higher than the traditional mix approach. Our approach is material to different properties that may be deduced for SNS clients, for example, age, vocation, and living arrangement. Since it is assumed that posted picture substance unmistakably reflects SNS client side interests and ways of life, our approach is appropriate for construing those qualities also. The primary problem is the SNS profile having their name, gender, age, a residence which is not openly to be had, however such facts is rather important for advertising. Text processor can handle the textual content facts that are accrued from person tweets in a SNS. As portrayed thus, we gathered two outcomes recovered by content and picture processors individually to upgrade the Twitter client gender orientation deduction.

Despite the fact that the gender orientation deduction exactness as of now came to 84.63 exclusively by the content classifier, we prevailing with regards to enhancing proficiency encourage by 0.48 pt. Since the picture preparing in our technique is totally free from the content handling, this joined the technique is material to the next gender orientation expectation strategies. Announced examinations about SNS client profile surmising focused on essential qualities, for example, gender orientation, age, vocation, local location, and so on. More advantageous qualities for promoting that specifically show client attributes are wanted to anticipate, for instance, interests and ways of life. Pictures in tweets are relied upon to incorporate signs about these profiles beside gender. The novel set of rules takes tweets as input and gives output as a gender opportunity score of the consumer. Since it is assumed that posted picture substance unmistakably reflects SNS client side interests and ways of life, our approach is appropriate for construing those qualities also. As a subject for future work, we will apply our consolidated strategy to different profile traits.

References

[1] J. Shetty Chandrakala, "Survey on Models to Investigate Data Center Performance and QoS in Cloud Computing Infrastructure", First International Conference on Recent Advances in Science & Engineering, **(2014)**.

[2] X. Ma, Y. Tsuboshita and N. Kato, "Gender estimation for SNS user profiling automatic image annotation", In Cross-media Analysis for Social Multimedia, **(2014)**.

[3] W. Liu and D. Ruths, "What's in a Name? Using First Names as Features for Gender Inference in Twitter", In Symposium on Analyzing Microtext, **(2013)**.

[4] Kazushi, Hattori, Hideki Asoh, Higashino "Twitter user profiling based on text and community mining for market analysis", **(2013)**.

[5] S. Sakaki, Y. Miura, M. Xiaojun, K. Hattori and T. Ohkuma, "Twitter user gender inference using combined analysis of text and image processing", **(2014)**, pp. 54-61.

[6] X. Ma, Y. Tsuboshita and N. Kato, "Gender estimation for SNS user profiling automatic image annotation", **(2014)**.

[7] A. Makazhanov and D. Refiei, "Predicting Political Preference of Twitter Users", In IEEE/ACM International Conference on Advances in Social Network and Mining, **(2013)**, pp. 298-305.

[8] A. Ulges, M. Koch and D. Borth, "Linking visual concept detection with viewer demographics", **(2012)**.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large scale visual recognition", **(2014)**.

[10] C. Sudhakar, A Sravani, N. Tirupathi Rao and D. Bhattacharyya, "By Text and Image Processing techniques to find SNS Profiler Gender Inference", Advanced Science and Technology Letters, vol. 147 (SMART DSC-2017), pp. 374-379, http://dx.doi.org/10.14257/astl.2017.147.53.

[11] A. Sravani, D.N.D. Harini, D. Lalitha Bhaskari"A Comparative Study of the Classification Algorithms Based on Feature Selection", ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol II, pp. 97-104.

[12] Yahoo! Japan. 2013. Yahoo Crowd Sourcing. Available: http://crowdsourcing.yahoo.co.jp/.

[13] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang and C.-J. Lin, "A library for large linear classification", **(2008)**.

[14] W. Liu and D. Ruths, "What's in a Name? Using First Names as Features for Gender Inference in Twitter", In **(2013)**.

Copyright ⓒ 2018 SERSC Australia